

On Bayesian quantile regression curve fitting via auxiliary variables

J.-L. Dortet-Bernadet

Institut de Recherche Mathématique Avancée, UMR 7501 CNRS
Université Louis Pasteur, Strasbourg, France

Y. Fan

School of Mathematics and Statistics
University of New South Wales, Sydney 2052, Australia

February 28, 2012

Abstract

Quantile regression has received increased attention in the statistics community in recent years. This article adapts an auxiliary variable method, commonly used in Bayesian variable selection for mean regression models, to the fitting of quantile regression curves. We focus on the fitting of regression splines, with unknown number and location of knots. We provide an efficient algorithm with Metropolis-Hastings updates whose tuning is fully automated. The method is tested on simulated and real examples and its extension to additive models is described. Finally we propose a simple postprocessing procedure to deal with the problem of the crossing of multiple separately estimated quantile curves.

Keywords: Quantile regression; Curve fitting; Gibbs sampling; Splines; Additive models; Automatic tuning; Noncrossing curves.

1 Introduction

Quantile regression has been recognized in recent years as a robust statistical procedure that offers a powerful alternative to the ordinary mean regression, especially when the data contains large outliers or when the response variable has a skewed or multimodal conditional distribution. Given a fixed probability p , $0 < p < 1$, let the model corresponding to the p -th quantile regression curve be given by

$$Y_i | x_1, \dots, x_n \sim f_p(x_i) + \epsilon_i, \quad i = 1, \dots, n$$

where $\epsilon_1, \dots, \epsilon_n$ are independent draws from a noise distribution whose p -th quantile is 0, i.e. $\mathbb{P}(\epsilon \leq 0) = p$. Under this model the p -th quantile of the conditional distribution

of Y given $\{X = x\}$ is given by some smooth function $f_p(x)$. If the distribution of the noise is left unspecified then the estimation of f_p is typically carried out by solving the minimization problem, for a given class \mathcal{F} of curves,

$$\arg \min_{f_p \in \mathcal{F}} \sum_{i=1}^n \rho_p(y_i - f(x_i)) \quad (1)$$

where the so-called "check function" $\rho_p(\cdot)$ is given by $\rho_p(\epsilon) = p\epsilon$ if $\epsilon \geq 0$ and $\rho_p(\epsilon) = (p-1)\epsilon$ otherwise (see Koenker and Bassett 1978). To define a likelihood function, one usually assumes that the noise distribution is an asymmetric Laplace distribution so that the maximum likelihood estimate corresponds to the solution of the minimization problem (see Koenker and Machado 1999). See e.g. Yu et al. (2003) or Koenker (2005) for a review on quantile regression and Geraci and Bottai (2007) for quantile regression with longitudinal data. References on Bayesian treatments of the subject include Tsionas (2003) for inference on a single quantile, Yu and Moyeed (2001) for quantile regression with a random walk Metropolis-Hastings algorithm and Yu (2002) for quantile regression with a reversible jump MCMC sampler (RJMCMC, Green 1995). More recently Yue and Rue (2011) considers additive mixed regression models and inference with either MCMC sampling or the integrated nested Laplace approximation (INLA, Rue et al. 2009) and Kozumi and Kobayashi (2011) proposed quantile regression with a Gibbs sampler.

In this article, we are interested in the case where the curve f_p is modeled with spline functions of a given degree, $P \geq 1$, so that,

$$f_p(x) = \alpha_0 + \sum_{j=1}^P \alpha_j x^j + \sum_{k=1}^K \eta_k (x - \gamma_k)_+^P \quad (2)$$

where $z_+ = \max(0, z)$ and where $\gamma_k, k = 1, \dots, K$ represent the locations of K knot points (see Hastie and Tibshirani 1990). Typically, the degree P is set to equal 3, since cubic splines are known to approximate locally smooth functions arbitrarily well. Chen and Yu (2009) provides a Bayesian inference on this model, where the number of knots and their location are automatically selected. Their method relies on a RJMCMC algorithm which, under the prior specifications they use, needs to compute an approximation of the ratio of marginal likelihoods. For fitting of quantile smoothing splines see Koenker et al. (1994) and He and Ng (1999) and for a Bayesian inference with natural cubic splines see Thompson et al. (2010).

We propose here an alternative strategy that avoids the use of the RJMCMC sampler which can often be difficult to tune (see Fan and Sisson 2011 for a review) and that does not rely on approximations to simplify computations. Recognising that a Bayesian variable selection technique (e.g. George and McCulloch 1993) can be used for inference on a curve (e.g. Smith and Kohn 1996, Fan et al. 2010) we use an auxiliary variable approach which makes possible, under appropriate prior specifications, a Metropolis-Hastings within Gibbs sampler. The proposed MCMC sampler is easy to implement and fully automated. In particular it incorporates an algorithm which automatically tunes the scaling parameters used in our Random-walk Metropolis-Hastings algorithm.

In Section 2 we present the model and the prior specifications, then we describe how inference is carried out with a MCMC sampler. We apply the method on several datasets in Section 3. In Section 4 we consider quantile curve regression for additive models. Finally, in Section 5, we discuss the problem of crossing quantile curves and propose a simple postprocessing procedure to reweight the MCMC samples from separately estimated quantile curves.

2 Quantile regression with splines

For some $0 < p < 1$, and given paired observations $(x_1, y_1), \dots, (x_n, y_n)$, we are interested in fitting the p -th quantile regression model

$$Y_i | x_1, \dots, x_n \sim f_p(x_i) + \epsilon_i, \quad i = 1, \dots, n \quad (3)$$

where $\epsilon_1, \dots, \epsilon_n$ are independent draws from the asymmetric Laplace distribution

$$d_{ALp(0, \sigma)}(\epsilon) = \frac{p(1-p)}{\sigma} \exp \left[-\frac{1}{\sigma} \rho_p(\epsilon) \right] \quad (4)$$

for an unknown scale parameter $\sigma > 0$. Under this model the p -th quantile of the conditional distribution of Y given $\{X = x\}$ is $f_p(x)$. The asymmetric Laplace distribution has been adopted in many papers, see for example Koenker and Machado (1999), Yu and Moyeed (2001), Tsionas (2003), Chen and Yu (2009), Yue and Rue (2011) or Kozumi and Kobayashi . Under the asymmetric Laplace distribution, given σ , the function f_p maximizing the likelihood corresponding to model (3) is also the solution of the minimization problem in Equation (1). The scale parameter σ that takes into account the variability of the observations is considered as a nuisance parameter.

We consider hereafter that the curve f_p is modeled with spline functions of a given degree $P > 0$, in the form of Equation (2). Under this representation, fitting the curve consists of estimating the number of knots K , the knot locations $\gamma_k, k = 1, \dots, K$, and the corresponding regression coefficients $\alpha_j, j = 0, \dots, P$ and $\eta_k, k = 1, \dots, K$. If $\gamma_k, k = 1, \dots, K_{max}$, where K_{max} represents the (known) maximum number of potential knots, model (3) can be written as the linear model

$$Y = X_\gamma \beta + \epsilon \quad (5)$$

where $Y = (y_1, \dots, y_n)'$, $\beta = (\alpha_0, \alpha_1, \dots, \alpha_P, \eta_1, \dots, \eta_{K_{max}})'$, $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$, with design matrix

$$X_\gamma = (\mathbf{1}_n, \mathbf{x}, \dots, \mathbf{x}^P, (\mathbf{x} - \mathbf{1}_n \gamma_1)_+^P, \dots, (\mathbf{x} - \mathbf{1}_n \gamma_{K_{max}})_+^P) \quad (6)$$

where $\mathbf{x} = (x_1, \dots, x_n)'$ and where $\mathbf{1}_n = (1, \dots, 1)'$ denotes the unit vector of size n .

2.1 The model and prior assumptions

We adopt an auxiliary variable approach for the spline regression model by introducing a vector of binary indicator variables $z_k, k = 1, \dots, K_{max}$,

$$z_k = \begin{cases} 1 & \text{if there is a knot point } \gamma_k \text{ in the interval } I_k \text{ and } \eta_k \neq 0 \\ 0 & \text{if there is no knot point in the interval } I_k \text{ and } \eta_k = 0 \end{cases}$$

where η_k denotes the spline coefficients in model (5), and the intervals I_k are defined on the range of the x_i 's. Each interval I_k contains at most one knot with unknown location γ_k . In practice, such intervals can be defined by either using prior information on regions where a knot is suspected or, in the absence of such prior information, an equal partition of the range may be adopted. We denote the vector $(\gamma_1, \dots, \gamma_{K_{max}})'$ by γ and consider the Uniform distributions on the interval as the prior distribution on γ . Each possible value for γ gives a model of the form (5). Let $X_{z, \gamma}$ denotes the matrix constructed with

the columns of X_γ corresponding to non-zero entries in z , and let $\beta_{z,\gamma}$ denotes the vector of corresponding regression coefficients.

A desirable feature of the asymmetric Laplace distribution is that it can be decomposed as a scale mixture of normals (see *e.g.* Tsionas 2003, Yue and Rue 2011 or Kozumi and Kobayashi 2011)

$$\begin{aligned}\epsilon|w &\sim \mathcal{N}\left(\frac{(1-2p)w}{p(1-p)}, \frac{2\sigma w}{p(1-p)}\right), \\ w &\sim \text{Exp}(1/\sigma),\end{aligned}$$

where $\text{Exp}(1/\sigma)$ denotes the exponential distribution with mean σ . If $w_i, i = 1, \dots, n$ denote the variable w associated with each ϵ_i , the conditional distribution of Y given W , the diagonal matrix with entries $w_i, i = 1, \dots, n$, is

$$f(Y|X_{z,\gamma}, \beta_{z,\gamma}, z, \sigma, \gamma, W) = \mathcal{N}\left(X_{z,\gamma}\beta_{z,\gamma} + \frac{(1-2p)}{p(1-p)}W\mathbf{1}_n, \frac{2\sigma}{p(1-p)}W\right). \quad (7)$$

Conditional on W we use the following decomposition of the joint prior distribution of the unknown parameters

$$\pi(\beta_{z,\gamma}, z, \sigma, \gamma|W) = \pi_{\beta_{z,\gamma}}(\beta_{z,\gamma}|z, \sigma, \gamma, W)\pi_\sigma(\sigma)\pi_z(z)\pi_\gamma(\gamma),$$

where we set

$$\pi_{\beta_{z,\gamma}}(\beta_{z,\gamma}|z, \sigma, \gamma, W) = \mathcal{N}\left(0, \frac{2\sigma}{p(1-p)}c(X'_{z,\gamma}W^{-1}X_{z,\gamma})^{-1}\right). \quad (8)$$

This conditional prior for $\beta_{z,\gamma}$, related to g -priors (Zellner 1986), has the advantage of conjugacy in the case of normal errors, in which case the regression and variance parameters can be analytically integrated out.

Different choices for the parameter c have been proposed in the literature for mean regression problems. The case $c = n$, where n is the sample size, corresponds to the unit information prior which was used by DiMatteo et al. (2001), a default choice that works well in practice in Bayesian variable selection problems with large sample sizes. Smith and Kohn (1996) recommend values of c in the range $10 \leq c \leq 1000$ for the problems they considered. Here including an adaptive scale parameter c , and treating it as another parameter was more satisfactory than using a fixed one. Thus we include a hyper-prior for c , following *e.g.* Leslie et al. (2007), we use a diffuse prior $\mathcal{IG}(1, 2n)$ with a mode at n

$$\pi(c) \propto c^{-2} \exp\{-2n/c\}.$$

See Liang et al. (2008) for more discussion about the choice of a prior distribution on the parameter c .

For the variance parameter, we use the standard uninformative prior $\pi_\sigma(\sigma) \propto 1/\sigma$. Finally, we need to define the prior distribution for z , we consider the decomposition of this prior given by

$$\pi_z(z) = \pi(z | |z|)\pi(|z|)$$

where $|z| = \sum_{k=1}^{K_{max}} z_k$ is the number of non-zero entries in z , *i.e.* the number of knots that are used in the corresponding model. We use for this term a Poisson distribution

with mean λ that is right-truncated at a specified maximum value, L . We assume also that, given this quantity, all possible configurations for z have equal probabilities, so that

$$\pi_z(z) \propto \frac{\lambda^{|z|}}{|z|!} I_{\{|z| \leq L\}}.$$

The parameters $\beta_{z,\gamma}$ and σ can be integrated out of the full joint posterior distribution $\pi(\beta_{z,\gamma}, z, \sigma, \gamma, W, c|Y)$ and we get

$$\pi(z, \gamma, W, c|Y) \propto \frac{\pi(c)\pi_z(z)\pi_\gamma(\gamma)}{\sqrt{\prod_{i=1}^n w_i(c+1)^{(|z|+P+1)/2}}} \left\{ \frac{p(1-p)}{4} S_{z,\gamma,W,c}(Y) + \sum_{i=1}^n w_i \right\}^{-3n/2} \quad (9)$$

where

$$S_{z,\gamma,W,c}(Y) = Y'_{(W)} W^{-1} Y_{(W)} - \frac{c}{c+1} Y'_{(W)} W^{-1} X_{z,\gamma} (X'_{z,\gamma} W^{-1} X_{z,\gamma})^{-1} X'_{z,\gamma} W^{-1} Y_{(W)}$$

and where

$$Y_{(W)} = Y - \frac{(1-2p)}{p(1-p)} W \mathbf{1}_n.$$

Details of the marginal posterior are given in Appendix A.

2.2 Inference on the posterior distribution

An MCMC sampler is used for the inference on the model. Based on the posterior distribution (9), for each t^{th} iteration of the MCMC update, $t = 1, \dots, T$, perform the following successive updates for z , γ , W and c :

- **Update z .** This update involves two types of moves; with probability 0.5 we propose an add/delete step, otherwise a swap step is proposed. Specifically, the two move steps involve
 - add/delete: randomly select a z_k and propose to change its value;
 - swap: randomly select two values z_i and z_j , and propose to exchange their values.

In both cases, proposed moves from current value z to proposed value z' are accepted with the usual Metropolis-Hastings acceptance probability

$$\alpha(z, z') = \min \left\{ 1, \frac{\pi(z', \gamma, W, c|Y) q(z', z)}{\pi(z, \gamma, W, c|Y) q(z, z')} \right\}$$

where $q(z, z')$ is the probability of proposing the new value z' given the current value z .

- **Update γ .** For each $k = 1, \dots, K_{max}$, we differentiate the cases when $z_k = 0$ and when $z_k = 1$:
 - if $z_k = 0$ then γ_k is updated according to its prior distribution, *i.e.* a Uniform distribution on I_k ;

- if $z_k = 1$, γ_k is updated to a new value γ'_k , according to the posterior distribution

$$\pi(\gamma_k | \gamma_{j \neq k}, z, Y, W, c) \propto \left\{ \frac{p(1-p)}{4} S_{z, \gamma, W, c}(Y) + \sum_{i=1}^n w_i \right\}^{-3n/2} \pi_\gamma(\gamma).$$

An independence Metropolis-Hastings step can be used for this last type of updating, using the prior on γ_k as a proposal, with the corresponding acceptance probability given by

$$\alpha(\gamma_k, \gamma'_k) = \min \left\{ 1, \frac{\pi(\gamma'_k | \gamma_{j \neq k}, z, Y, W, c)}{\pi(\gamma_k | \gamma_{j \neq k}, z, Y, W, c)} \right\}.$$

- **Update W .** Each $w_i, i = 1, \dots, n$ has conditional posterior distribution

$$\pi(w_i | w_{i \neq j}, \gamma, z, c, Y) \propto \frac{1}{\prod_{i=1}^n \sqrt{w_i}} \left\{ \frac{p(1-p)}{4} S_{z, \gamma, W, c}(Y) + \sum_{i=1}^n w_i \right\}^{-3n/2}.$$

We use a Random-walk Metropolis-Hastings proposal to update each w_i . We consider as proposal distribution a normal distribution $q(w_i, \cdot) = N(w_i, \sigma_i^2)$ with mean w_i and variance σ_i^2 . We sample $w'_i \sim q(w_i, \cdot)$, then the proposed value w'_i is accepted with probability

$$\alpha(w_i, w'_i) = \min \left\{ 1, \frac{\pi(w'_i | w_{i \neq j}, \gamma, z, c, Y)}{\pi(w_i | w_{i \neq j}, \gamma, z, c, Y)} \right\}.$$

The tuning parameters $\sigma_i^2, i = 1, \dots, n$ are optimally obtained automatically, prior to starting the main part of MCMC, see Appendix B.

- **Update c .** The parameter c has conditional distribution

$$\pi(c | z, \gamma, W, Y) \propto \frac{\pi(c)}{(c+1)^{(|z|+P+1)/2}} \left\{ \frac{p(1-p)}{4} S_{z, \gamma, W, c}(Y) + \sum_{i=1}^n w_i \right\}^{-3n/2}.$$

We use a Random-walk Metropolis-Hastings proposal to update c . We sample $c' \sim q(c, \cdot) = N(c, \sigma_*^2)$ then accept the proposed value with acceptance probability

$$\alpha(c, c') = \min \left\{ 1, \frac{\pi(c' | z, \gamma, W, Y)}{\pi(c | z, \gamma, W, Y)} \right\}.$$

The tuning parameter σ_*^2 is also obtained via the algorithm in Appendix B.

Note that when the sample size n is large, the number of parameters in the Update W step becomes large and correspondingly manual tuning of the scale parameters σ_i^2 in the Gaussian Random-Walk Metropolis-Hastings sampler becomes infeasible. One strategy to automate the sampler is to use a slice sampler (see Neal 2003). But the additional evaluations of the posterior function makes this algorithm much more computationally intensive. In this article we use the algorithm of Garthwaite et al. (2010) that automatically tunes the scaling parameters σ_i^2 and obtains an optimal over all acceptance rate of $p^* = 0.44$ (Roberts and Rosenthal 2001) for these univariate updates. See Appendix B for a description of the algorithm used for tuning.

Once a converged MCMC sample $\{(z^{(t)}, \gamma^{(t)}, W^{(t)}, c^{(t)})\}_{t=1, \dots, T}$ is obtained it is possible to estimate the curve $f_p(x)$ by a Bayesian model averaging approach (BMA). The posterior expectation for β given z, γ, W and c is

$$\mathbb{E}(\beta_{z,\gamma}|z, \gamma, W, Y, c) = \frac{c}{c+1} (X'_{z,\gamma} W^{-1} X_{z,\gamma})^{-1} X'_{z,\gamma} W^{-1} Y_{(W)}. \quad (10)$$

Thus an estimate for $f_p(x)$ can be obtained by

$$\hat{f}_p^{BMA}(x) = \frac{1}{T} \sum_{t=1}^T X_{z^t, \gamma^t} \frac{c^{(t)}}{c^{(t)} + 1} (X'_{z^t, \gamma^t} (W^t)^{-1} X_{z^t, \gamma^t})^{-1} X'_{z^t, \gamma^t} (W^t)^{-1} Y_{(W^t)}.$$

Another possibility to estimate the curve $f_p(x)$ is to use the maximum a posteriori (MAP) estimate for (z, γ, W, c)

$$(\hat{z}, \hat{\gamma}, \hat{W}, \hat{c}) = \underset{1 \leq t \leq T}{\operatorname{argmax}} \pi(z^{(t)}, \gamma^{(t)}, W^{(t)}, c^{(t)} | Y),$$

then calculate the corresponding curve estimate via

$$\hat{f}_p^{MAP}(x) = \frac{\hat{c}}{\hat{c} + 1} X_{\hat{z}, \hat{\gamma}} (X'_{\hat{z}, \hat{\gamma}} \hat{W}^{-1} X_{\hat{z}, \hat{\gamma}})^{-1} X'_{\hat{z}, \hat{\gamma}} \hat{W}^{-1} Y_{(\hat{W})}.$$

3 Examples

3.1 Simulation studies

We carry out simulations to compare the use of the method described in this paper with the method COBS proposed by He and Ng (1999). COBS estimates both constrained and unconstrained quantile curves using B-spline smoothing and is available as an R package. Here we use the unconstrained case as a fully automated procedure, where both the smoothing parameter and the selection of knots is carried out according to either the AIC or the BIC criterion.

We consider simulated datasets that correspond to the three examples described below, these examples are adapted from some well known examples in the curve fitting literature, see e.g. Smith and Kohn (1996), Denison et al. (1998) and DiMatteo et al. (2001).

Example 1: Here the curve takes the form

$$f(x) = \phi(x, 0.15, 0.05^2)/4 + \phi(x, 0.6, 0.2^2)/4, \quad x \in [0, 1],$$

where $\phi(x, \mu, \sigma^2)$ denotes the value at x of the normal density with mean μ and variance σ^2 . $n = 200$ data points x are sampled from the Uniform distribution $U(0, 1)$. The noise ϵ is added to the data, they corresponds to a Gamma distribution $\mathcal{G}a(1, 4)$ with shape parameter 1 and rate parameter 4 that is translated by -0.175 (so that the median of this noise distribution is approximatively 0).

Example 2: Here the curve takes the form

$$f(x) = \sin(2x) + 2 \exp(-16x^2), \quad x \in [-2, 2].$$

and is evaluated at $n = 201$ regularly spaced grid points. This function is first rescaled so that the support is on the unit interval. The noise ϵ added to the data is simulated in the same way as in the first example.

Example 3: In this example the curve is given by,

$$f(x) = \sin(x) + 2 \exp(-30x^2), \quad x \in [-2, 2],$$

and the data points x correspond to $n = 201$ regularly spaced grid points. As in the previous example the function is rescaled on the unit interval for x and the same distribution for noise ϵ is used for the data.

To compare the different methods we use the mean squared error (MSE) as a measure of goodness of fit, given by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \{\hat{f}(x_i) - f(x_i)\}^2$$

where f is the true median regression function and \hat{f} is the estimated function. Since the COBS algorithm computes the median curve with quadratic (or linear) splines we consider hereafter the case $P = 2$.

For the prior specifications of each example, we set $\lambda = 3$ and $L = 10$ for the truncated Poisson prior. Results are largely insensitive to values of λ around this range and the maximum number of knots allowed L is chosen to be large enough to not affect the simulation results here. For these examples we consider the situation where there is no prior information on the knot locations and chose the intervals I_k to correspond to the ranges given by every n_x sorted x values. We found that $n_x = 5$ was sufficient to provide a good fit in each of the three examples. We use a B-spline basis to formulate the $X_{z,\gamma}$ matrix, as in DiMatteo et al. (2001), to avoid numerical instability (see e.g. Ruppert et al. 2003).

The computation of all three examples started with an arbitrary set of initial values generated from the prior distributions. We first ran the algorithm 500 iterations for adaptive tuning then fixing the scaling parameters of the Random-Walk Metropolis Hastings algorithm at the final value of the tuning run, we then ran a burn-in of 500 iterations, followed by 1,500 recorded iterations. Each iteration involves an update of 20 z update steps for each γ update step. To assess convergence, we monitored the trace plots of each model parameters as well as posterior values. We also ran much longer chains of 10,000 iterations and found the results to be similar in terms of MSE calculations. See Figure 1 for the fitted functions of the three examples using our method with the BMA estimate and with the MAP estimate.

For each of the three examples the BMA, MAP and COBS (with the AIC or the BIC criterion) estimates are calculated over 50 randomly generated datasets. The mean and standard deviation of the MSEs are presented in Table 1, the corresponding boxplots are given in Figure 2. On the whole the method presented in this paper performs well compared to COBS, especially on the datasets corresponding to Example 3. On the three types of datasets that are considered here, the BMA estimates seem to be more accurate than the MAP estimates.

3.2 Motorcycle data set

We consider a reference dataset, the motorcycle data, studied in the context of quantile regression for example in Koenker (2005) or in Chen and Yu (2009). These data are analyzed in Silverman (1985) and contain experimental measurements of the acceleration of the head of a test dummy (expressed in g , acceleration due to gravity) as a function of

	BMA	MAP	COBS AIC	COBS BIC
Example 1	0.0032 (0.0017)	0.0055 (0.0021)	0.0052 (0.0033)	0.0075 (0.0069)
Example 2	0.0040 (0.0025)	0.0067 (0.0038)	0.0060 (0.0029)	0.0065 (0.0026)
Example 3	0.0036 (0.0018)	0.0056 (0.0025)	0.0084 (0.0028)	0.0139 (0.0044)

Table 1: Mean MSEs with estimated standard errors in brackets based on 50 samples obtained using the Bayesian model averaging (BMA), the maximum a posteriori (MAP) and the COBS algorithm (with the AIC criterion and with the BIC criterion).

time in the first moments after an impact (the time is expressed in *ms*). The dataset is challenging for quantile regression as the values and the variability of the response vary dramatically with the independent variable.

We fit to these data the quantile regression curves corresponding to $p = 0.25, 0.5$ and 0.75 . The prior settings are essentially the same as the ones already described in the simulation studies, except here we set $\lambda = 5$ and $L = 15$ for the truncated Poisson prior. For the MCMC computation of the curves we started with an arbitrary set of initial values generated from the prior distributions. Again we used the first 500 iterations for adaptive tuning then we ran a burn-in of 500 iterations followed by 3,500 recorded iterations, where each iteration involves an update of 20 z update steps for each γ update step.

We give in Figure 3 the quantile curves corresponding to linear splines $P = 1$. The results appear quite satisfactory as the quantile curves are not crossing each other, even in the region beyond 50 millisecond where the data are sparse. The changes in the variability of the acceleration over time has been captured well by the fitted conditional quantile curves, as they are very close to each other for the first few milliseconds then diverge after the crash.

4 Quantile regression for additive models

4.1 Introduction

When several potential predictors for the response variable are of interest, a standard procedure to avoid the so-called “curse of dimensionality” is to use an additive model (Hastie and Tibshirani 1990) where the response is modeled as a sum of functions of the predictors. In the context of quantile regression, if Y denotes the real-valued response variable and if now $\mathbf{X} = (X^1, \dots, X^d)$ denotes a vector of d predictors, the p -th quantile of the conditional distribution of Y given $\{\mathbf{X} = \mathbf{x}\}$ is modeled as

$$f_p(\mathbf{x}) = \sum_{j=1}^d f_p^j(x^j). \quad (11)$$

See Yu and Lu (2004) for an inference on the additive quantile regression model by a kernel-weighted local linear fitting and see Yue and Rue (2011) for a Bayesian inference with either a MCMC algorithm or using INLA.

If we use spline functions to model the different curves $f_p^1(x^1), \dots, f_p^d(x^d)$ it is still possible to use the linear model (5) with the difference that the design matrix X_γ is now made up of the columns of the individual design matrices corresponding to (6), with a single intercept term for identifiability. Thus inference on the additive quantile regression model can be performed via the same methodology and algorithm described in the previous sections. We consider below the study of a real dataset that involves additive quantile regression.

4.2 Analysis of the Boston housing dataset

We revisit the so-called Boston house price data available in the R package MASS. This dataset has been originally studied in Harrison and Rubinfeld (1978). The full dataset consists of the median value of owner-occupied homes in 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970 along with 13 various sociodemographic variables. This dataset has been analyzed in many statistical papers including Opsomer and Ruppert (1998) who used an additive model for mean regression, and Yu and Lu (2004), who proposed an additive quantile regression model by a kernel-weighted local linear fitting. As in these two references we consider the median values of the owner-occupied homes (in \$1000s) as the dependent variable and four covariates given by

RM = average number of rooms per house in the area,
TAX = full property tax rate (\$/\$10,000),
PTRATIO = pupil/teacher ratio by town school district,
LSTAT = the percentage of the population having lower economic status in the area.

As noticed in Yu and Lu (2004) these data are suitable for a quantile regression analysis since the response is a median price in a given area and the variables RM and LSTAT are highly skewed. More precisely we consider the additive model where the p -th quantile of the conditional distribution of the response is given by

$$f_p(\mathbf{x}) = \alpha_0 + f_p^1(\text{RM}) + f_p^2(\log(\text{TAX})) + f_p^3(\text{PTRATIO}) + f_p^4(\log(\text{LSTAT})).$$

We fit to these data the p -th quantile regression curves corresponding to cubic splines ($P = 3$) at the quantile levels $p = 0.25, 0.5$ and 0.75 . For the prior settings we took $\lambda = 5$ and $L = 8$ for the truncated Poisson prior. For each predictor we set the intervals I_k to be 10 equally sized partition sets over the range of the variable. Excluding the possibility of knots in the first and the last intervals, we get $K_{max} = 8$ for each variable. For the MCMC computation of the curves we started with a random set of initial values generated from the prior distributions. We first ran the algorithm 500 iterations for adaptive tuning then we ran a burn-in of 500 iterations, followed by 4,000 recorded iterations, where each iteration involves an update of 20 z update steps for each γ update step. We present in Figure 4 the different estimated curves. We plotted on the same graphs the datapoints corresponding to the original data minus the effect of all the other variables and the constant term. The fact that the values of $\log(\text{TAX})$ are not well dispersed over their range and the presence of a few outliers in the dataset did not seem to be a problem for our method.

Our results appear consistent with the results provided in the quoted previous analyses. Briefly, the variables RM and LSTAT appear as the most important covariates. If the contribution of LSTAT look similar for the three quantiles levels, the contribution of RM looks slightly more important for the upper quantile level $p = 0.75$. The variable TAX has a contribution relatively more important for the lower quantile level $p = 0.25$. Finally the Figure 4 suggests a linear contribution of the variable PTRATIO, especially for $p = 0.5$ and for $p = 0.75$.

5 Noncrossing quantile regression curves

One known problem when using quantile regression for multiple percentiles is that the quantile curves that are estimated separately can cross, which is impossible. See for example the Figure 5 (a) where, partly due to the relatively small size of the dataset and the complex conditional distribution of the response variable, the two estimated quantile curves for $p = 0.2$ and $p = 0.3$ are crossing around the value $x = 0.6$.

The treatment of noncrossing quantile regression curves is difficult and several attempts to circumvent this problem have been proposed in different settings, see *e.g.* the references in Yu et al. (2003) and in Koenker (2005) or, for a more recent development in this area, see *e.g.* Reich et al. (2011). In particular, Bondell et al. (2010) proposed a solution to this problem by considering a generalization of the criterion (1) to the case of simultaneous inference on several quantile curves. For clarity we suppose hereafter that we are interested in the fitting of two quantile curves corresponding to quantile levels p_1 and p_2 , with $p_1 < p_2$. Bondell et al. (2010) gave a solution to the minimization under the constraint $f_{p_1}(\cdot) < f_{p_2}(\cdot)$ of the expression

$$\sum_{j=1}^2 \left\{ \sum_{i=1}^n \rho_{p_j}(y_i - f_{p_j}(x_i)) \right\} \quad (12)$$

plus a penalty term corresponding to smoothing. An alternative approach described in Dunson and Taylor (2005) uses a so-called “substitution likelihood” that does not correspond to the distribution of the data given the unknown curves but yields a valid uncertainty. The substitution likelihood that they considered corresponds to the multinomial weights

$$s(f_{p_1}, f_{p_2}|Y) = \frac{n!}{u_1!u_2!u_3!} p_1^{u_1} (p_2 - p_1)^{u_2} (1 - p_2)^{u_3} I_{\{f_{p_1} < f_{p_2}\}} \quad (13)$$

where u_1 represents the number of datapoints below the curve f_1 , where u_2 represents the number of datapoints between the two curves and where u_3 is the number of datapoints above the curve f_2 . They gave conditions on the prior for the “pseudo-posterior” $\pi(f_{p_1}, f_{p_2}|Y) \propto s(f_{p_1}, f_{p_2}|Y)\pi(f_{p_1}, f_{p_2})$ to be proper and proposed a MCMC algorithm for (pseudo-)posterior computation in the case of linear quantiles.

Here we propose a new method to postprocess the MCMC samples obtained from separate quantile regression curve fitting. We denote by $\theta_p = (\beta_{z,\gamma}, \sigma, z, \gamma, W, c)$ the full set of unknown parameters for the p -th quantile regression curve model (7). Let $\theta_{p_1}, \theta_{p_2}$, be the parameters corresponding to the quantile regression curve for the quantile levels p_1 and p_2 respectively, $p_1 < p_2$. We consider a new substitution likelihood of the form

$$s(\theta_{p_1}, \theta_{p_2}|Y) = L(\theta_{p_1}|Y)L(\theta_{p_2}|Y)I_{\{f_{p_1}(x|\theta_{p_1}) < f_{p_2}(x|\theta_{p_2})\}} \quad (14)$$

where $L(\theta_{p_1}|Y)$ and $L(\theta_{p_2}|Y)$ denotes the two likelihood functions for quantile levels p_1 and p_2 given by the conditional distribution (7). The indicator function takes the value one if $f_{p_1}(x|\theta_{p_1}) < f_{p_2}(x|\theta_{p_2})$ for all x and zero otherwise, here the function $f_p(x|\theta_p)$ is evaluated according to Equation (2) with parameters θ_p . It is not hard to see that the maximizer of this substitution likelihood is the maximizer of (12). Moreover, if we take independent priors $\pi(\theta_{p_1})$ and $\pi(\theta_{p_2})$ on the two sets of parameters, then the corresponding quasi-posterior is simply

$$\begin{aligned}\pi(\theta_{p_1}, \theta_{p_2}|Y) &\propto s(\theta_{p_1}, \theta_{p_2}|Y)\pi(\theta_{p_1})\pi(\theta_{p_2}), \\ &\propto \pi(\theta_{p_1}|Y)\pi(\theta_{p_2}|Y)I_{\{f_{p_1}(x|\theta_{p_1}) < f_{p_2}(x|\theta_{p_2})\}}.\end{aligned}\quad (15)$$

Given samples from the distribution $\pi(\theta_{p_1}|Y) \otimes \pi(\theta_{p_2}|Y)$ an importance sampling argument can be used to reweight the samples according to this quasi-posterior.

In practice, MCMC samples obtained from separate posterior explorations of $\pi(\theta_{p_1}|Y)$ and of $\pi(\theta_{p_2}|Y)$ can be combined to form the new estimate of the the curve $f_{p_1}(x)$ by

$$\mathbb{E}_{\pi(\theta_{p_1}, \theta_{p_2}|Y)}[f_{p_1}(x|\theta_{p_1})] \approx \frac{\sum_t f_{p_1}(x|\theta_{p_1}^t) I_{\{f_{p_1}(x|\theta_{p_1}^t) < f_{p_2}(x|\theta_{p_2}^t)\}}}{\sum_t I_{\{f_{p_1}(x|\theta_{p_1}^t) < f_{p_2}(x|\theta_{p_2}^t)\}}}.\quad (16)$$

When the constraint above excludes too many samples this estimator will be unreliable, in this case more MCMC samples will be required. A computationally cheap way to obtain more samples is to consider all combinations of the two MCMC samples.

Figure 5 (b) shows the corrected curves from the estimator in (16), using all the possible combinations of the two MCMC samples, each of size 2,000. To evaluate the curves we use here the plug-in estimator (10) for $\beta_{z,\gamma}$. Finally the constraint on the curves is checked at every observed values of x .

The strength of the above approach is that it is very easy to apply, and can be used on any posterior samples from separate quantile curves. An obvious draw back is that in some cases, when for example p_1 and p_2 are very close, the number of samples satisfying the constraint can be extremely low.

6 Conclusion

In this article, we have provided a procedure for Bayesian inference on quantile curve fitting. We focused on the use of regression splines with unknown number of knots and location to obtain smooth curves. We have seen that, within an auxiliary variable framework, a scale mixture of normals representation for the asymmetric Laplace distribution together with appropriate prior specifications makes it possible to integrate out the regression and the variance parameters analytically. This facilitates a simple Metropolis-Hastings within Gibbs sampler for simulation from the posterior distribution of interest. The proposed algorithm is fully automated with the inclusion of an automatic tuning step, which optimally tunes the Random-Walk Metropolis-Hastings scaling parameters. We have shown that our method performs well on several types of datasets. We have also shown that the proposed framework can be trivially extended to inference on additive models. Finally we have proposed and discussed a simple and general procedure that postprocesses MCMC samples to obtain noncrossing quantile regression curves.

References

- Bondell, H. D., B. J. Reich, and H. Wang (2010). Noncrossing quantile regression curve estimation. *Biometrika* 97(4), 825–838.
- Chen, C. and K. Yu (2009). Automatic Bayesian quantile regression curve fitting. *Statistics and Computing* 19, 271–281.
- Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998). Automatic Bayesian curve fitting. *Journal of Royal Statistical Society, Series B* 60, 330 – 350.
- DiMatteo, I., C. R. Genovese, and R. E. Kass (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* 88(4), 1055–1071.
- Dunson, D. B. and J. A. Taylor (2005). Approximate Bayesian inference for quantiles. *J. Nonparametr. Stat.* 17(3), 385–400.
- Fan, Y., J.-L. Dortet-Bernadet, and S. A. Sisson (2010). On Bayesian curve fitting via auxiliary variables. *J. Comput. Graph. Statist.* 19(3), 626–644.
- Fan, Y. and S. A. Sisson (2011). *Handbook of Markov Chain Monte Carlo*, Chapter Reversible Jump Markov chain Monte Carlo. Chapman and Hall/CRC Press.
- Garthwaite, P. H., Y. Fan, and S. A. Sisson (2010). Adaptive optimal scaling of metropolis-hastings algorithms using the robbins-monro process. Technical report, University of New South Wales.
- George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of American Statistical Association* 88, 881 – 889.
- Geraci, M. and M. Bottai (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* 8(1), 140–154.
- Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Harrison, D. J. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5(1), 81–102.
- Hastie, T. J. and R. J. Tibshirani (1990). *Generalised additive models*. Chapman and Hall, London.
- He, X. and P. Ng (1999). Cobs: Qualitatively constrained smoothing via linear programming. *Computational Statistics* 14(3), 315–337.
- Koenker, R. (2005). *Quantile regression*, Volume 38 of *Econometric Society Monographs*. Cambridge: Cambridge University Press.
- Koenker, R. and J. Bassett, Gilbert (1978). Regression quantiles. *Econometrica* 46(1), 33–50.
- Koenker, R. and J. A. F. Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* 94(448), pp. 1296–1310.
- Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika* 81(4), 673–680.
- Kozumi, H. and G. Kobayashi (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation* 81(11), 1565–1578.
- Leslie, D., R. Kohn, and D. Nott (2007). A general approach to heteroscedastic linear regression. *Statistics and Computing* 17, 131–146.

- Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008, March). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* 103, 410–423.
- Neal, R. M. (2003). Slice sampling. *Annals of Statistics* 31(3), 705 – 767.
- Opsomer, J. D. and D. Ruppert (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association* 93(442), pp. 605–619.
- Reich, B. J., M. Fuentes, and D. B. Dunson (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association* 106(493), 6–20.
- Roberts, G. O. and J. S. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* 16, 351–367.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(2), 319–392.
- Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*. Cambridge University Press.
- Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)* 47(1), pp. 1–52.
- Smith, M. and R. Kohn (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* 75, 317–343.
- Thompson, P., Y. Cai, R. Moyeed, D. Reeve, and J. Stander (2010). Bayesian nonparametric quantile regression using splines. *Computational Statistics and Data Analysis* 54(4), 1138 – 1150.
- Tsionas, E. G. (2003). Bayesian quantile inference. *J. Stat. Comput. Simul.* 73(9), 659–674.
- Yu, K. (2002). Quantile regression using RJMCMC algorithm. *Comput. Statist. Data Anal.* 40(2), 303–315.
- Yu, K. and Z. Lu (2004). Local linear additive quantile regression. *Scandinavian Journal of Statistics* 31(3), pp. 333–346.
- Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *The Statistician* 52(3), 331–350.
- Yu, K. and R. A. Moyeed (2001). Bayesian quantile regression. *Statist. Probab. Lett.* 54(4), 437–447.
- Yue, Y. R. and H. Rue (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics and Data Analysis* 55(1), 84 – 96.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g -prior distributions. In *Bayesian inference and decision techniques*, Volume 6 of *Stud. Bayesian Econometrics Statist.*, pp. 233–243. Amsterdam: North-Holland.

Appendix A

The marginal posterior

The full joint posterior distribution of the parameters is

$$\begin{aligned}\pi(\beta_{z,\gamma}, z, \sigma, \gamma, W, c|Y) &\propto f(Y|X_{z,\gamma}, \beta_{z,\gamma}, z, \sigma, \gamma, W)\pi(W|\sigma)\pi(\beta_{z,\gamma}, z, \sigma, \gamma|W)\pi(c), \\ &\propto f(Y|X_{z,\gamma}, \beta_{z,\gamma}, z, \sigma, \gamma, W)\pi(W|\sigma)\pi_{\beta_{z,\gamma}}(\beta_{z,\gamma}|z, \sigma, \gamma, W) \\ &\quad \times \pi_{\sigma}(\sigma)\pi_z(z)\pi_{\gamma}(\gamma)\pi(c).\end{aligned}$$

With $f(Y|X_{z,\gamma}, \beta_{z,\gamma}, z, \sigma, \gamma, W)$ and $\pi_{\beta_{z,\gamma}}(\beta_{z,\gamma}|z, \sigma, \gamma, W)$ given by the two Gaussian distributions (7) and (8) the parameter $\beta_{z,\gamma}$ is easily integrated out using classical results of Bayesian linear models. We get

$$\pi(z, \sigma, \gamma, W, c|Y) \propto \left(\frac{1}{c+1}\right)^{\frac{|z|+P+1}{2}} \frac{\pi_z(z)\pi_{\gamma}(\gamma)}{\sqrt{\pi_{i=1}^n w_i}} \left(\frac{1}{\sigma}\right)^{1+\frac{3n}{2}} e^{-\frac{1}{\sigma}\left\{\frac{p(1-p)}{4}S_{z,\gamma,W}(Y) + \sum_{i=1}^n w_i\right\}} \pi(c)$$

where

$$S_{z,\gamma,W}(Y) = Y'_{(W)}W^{-1}Y_{(W)} - \frac{c}{c+1}Y'_{(W)}W^{-1}X_{z,\gamma}(X'_{z,\gamma}W^{-1}X_{z,\gamma})^{-1}X'_{z,\gamma}W^{-1}Y_{(W)}$$

with

$$Y_{(W)} = Y - \frac{(1-2p)}{p(1-p)}W\mathbf{1}_n.$$

Then the parameter σ can be also integrated out and we get

$$\pi(z, \gamma, W, c|Y) \propto \left(\frac{1}{c+1}\right)^{\frac{|z|+P+1}{2}} \frac{\pi_z(z)\pi_{\gamma}(\gamma)}{\sqrt{\pi_{i=1}^n w_i}} \left(\frac{1}{\frac{p(1-p)}{4}S_{z,\gamma,W,c}(Y) + \sum_{i=1}^n w_i}\right)^{\frac{3n}{2}} \pi(c).$$

Appendix B

Automatic tuning algorithm

Here we provide the algorithm to optimally search for the tuning parameters $\sigma_i^2, i = 1, \dots, n$, and σ_*^2 . The algorithm runs within the main MCMC algorithm given in Section 2.2. Tuning will only apply to the Update W and Update c steps. For the update of each of the parameters $w_i, i = 1, \dots, n$ and c do:

Initialisation: For the iteration $t = 1$ of the algorithm initialise the scaling parameter $\sigma^* = \sigma^1 = 1$, where σ^1 corresponds to σ_i and σ_* when updating the parameters $w_i, i = 1, \dots, n$ and c respectively. Set $p^* = 0.44$ and initialise $j = 0$. The value of p^* corresponds to the optimal acceptance probability for a univariate Random-Walk Metropolis-Hastings algorithm (Roberts and Rosenthal 2001).

Tuning: Set $j = j + 1$; update the parameters according to either Update W or update c , and obtain the corresponding acceptance probability α as in Section 2.2.

update scaling: if $j < 20$ set $\sigma^{t+1} = \sigma^t$, else set

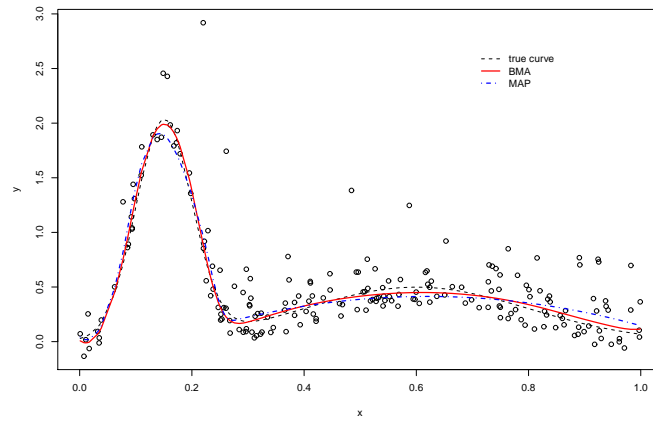
$$\sigma^{t+1} = \begin{cases} \sigma^t + \kappa(1 - p^*)/j & \text{if } U < \alpha \\ \sigma^t - \kappa p^*/j & \text{if } U > \alpha \end{cases}$$

where $\kappa = \sigma^t / \{p^*(1 - p^*)\}$ and $U \sim U(0, 1)$.

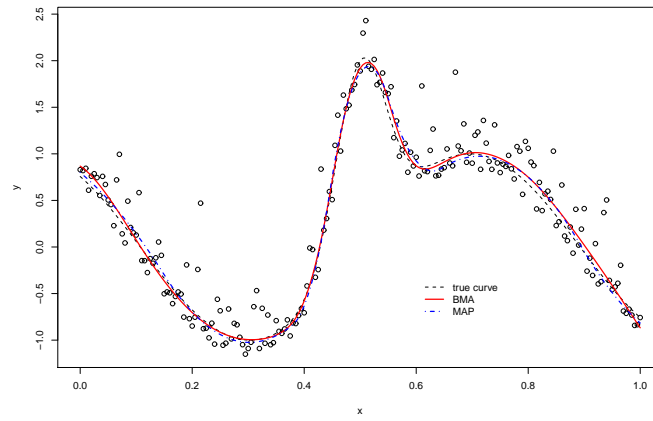
restart the algorithm: If $t < 100$, and either $\sigma^{t+1} > 3\sigma^*$ or $\sigma^{t+1} < \sigma^*/3$, restart the algorithm, setting $\sigma^* = \sigma^{t+1}$ and $j = 0$. Note we do not restart the algorithm again if the total number of restarts exceeds 5.

Increment loop: Set $t = t + 1$. Go back to the beginning unless t exceeds some pre-specified number of iterations $nTune$.

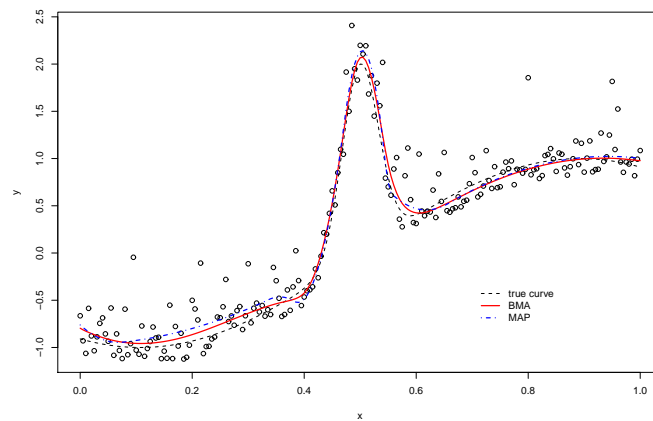
It is easy to monitor the changes in σ^t in order to determine the number of tuning iterations $nTune$ to achieve stability. In practice, we run the first $nTune$ iterations of the algorithm in Section 2.2 with automatic tuning, and then start the main part of MCMC as usual with the scaling parameters fixed at the value σ^{nTune} .



(a) Example 1

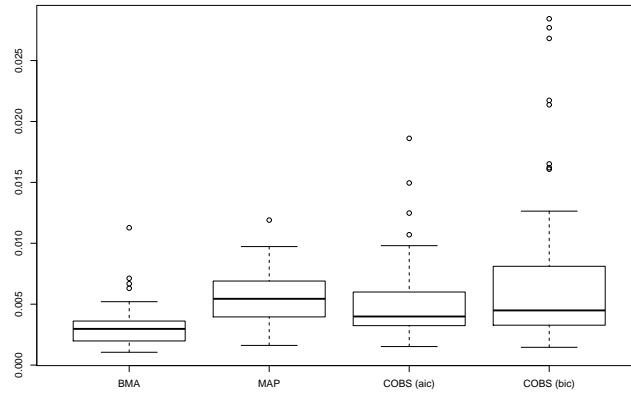


(b) Example 2

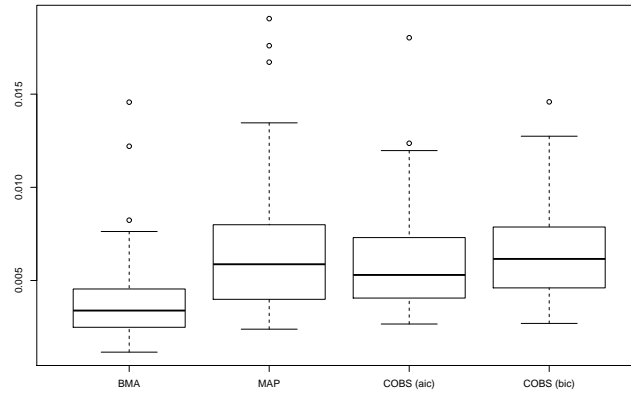


(c) Example 3

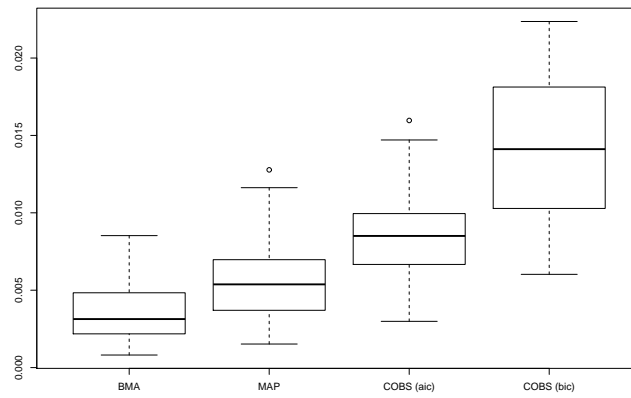
Figure 1: Estimated curves for the three simulated examples.



(a) Example 1



(b) Example 2



(c) Example 3

Figure 2: Boxplots for the MSEs corresponding to the three simulated examples.

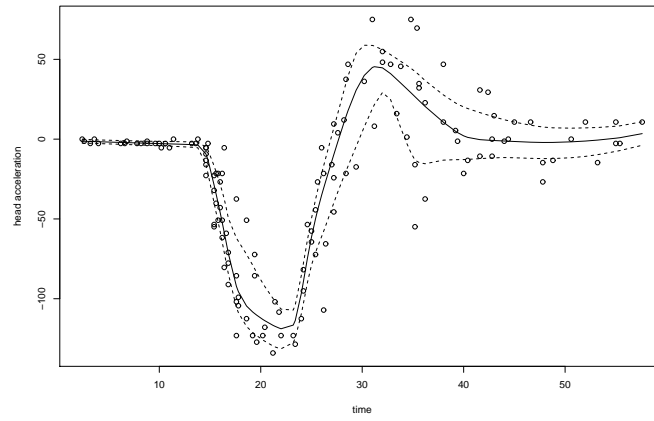


Figure 3: Motor cycle data set ; estimated quartile regression curves by BA approach for $P = 1$ (p=0.5: solid line ; p=0.25 and p=0.75: dotted lines)

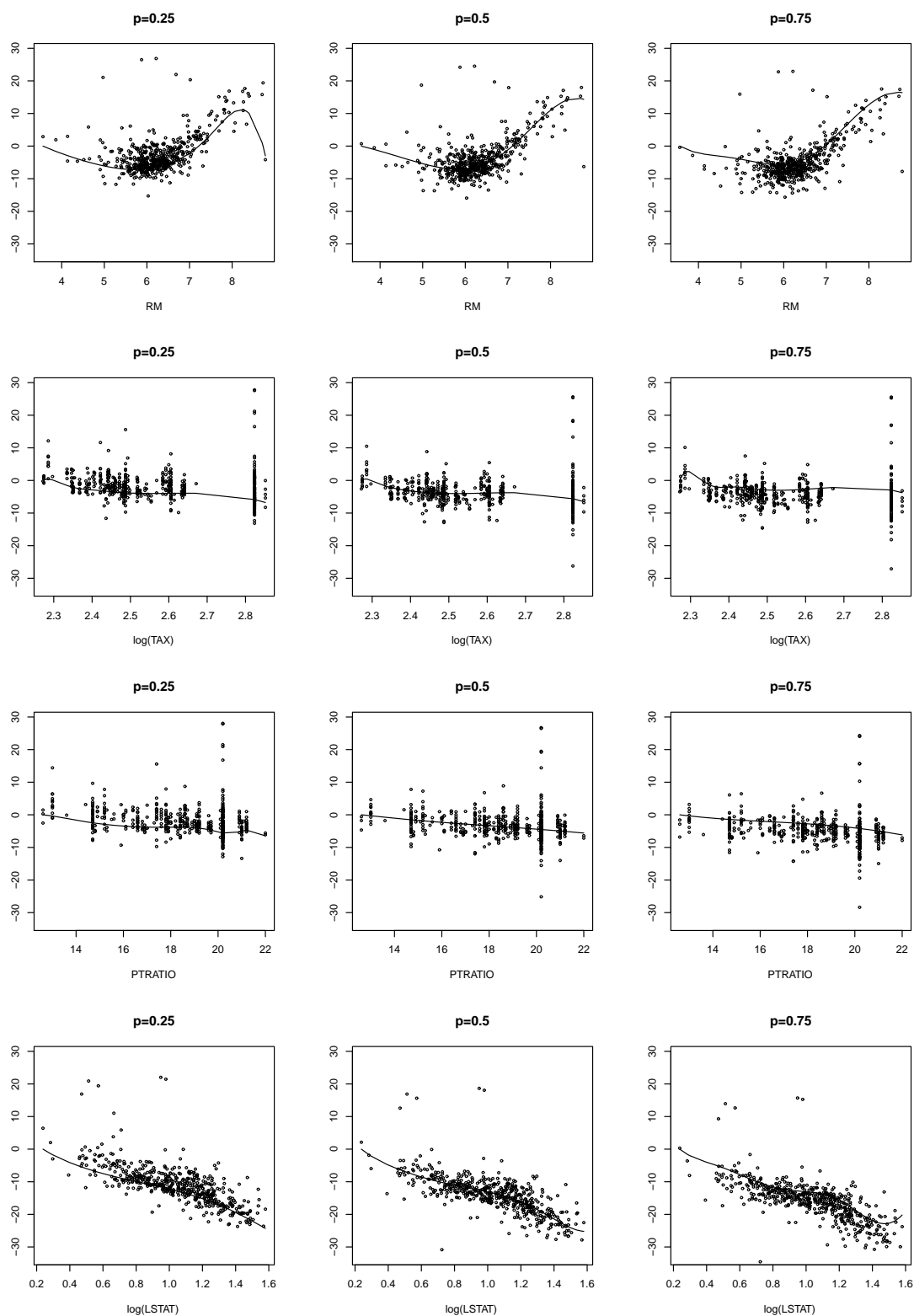
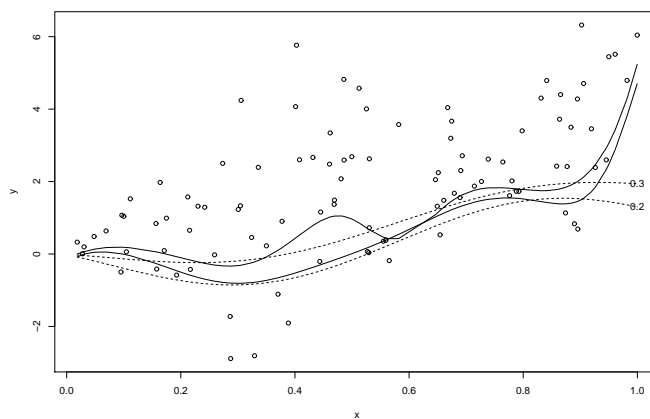
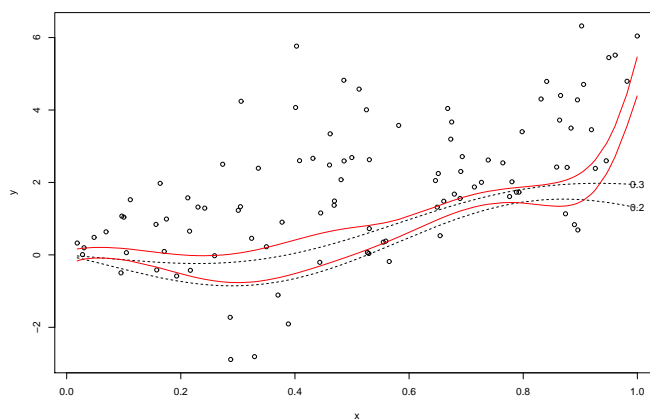


Figure 4: Boston housing dataset ; fitted quantile curves $p = 0.25, 0.5, 0.75$, $P = 3$, for the four variables that have been considered. One each figure the datapoints represented correspond to the original data minus the effect of all the other variables (and the constant term).



(a) Initially estimated curves



(b) Corrected curves

Figure 5: Curve crossing example. The dotted lines represent the true quantile curves for $p = 0.2$ and $p = 0.3$. The solid lines represent (a) the quantiles curves that have been estimated separately (b) the corrected estimated quantile curves with respect to the new substitution likelihood.

